Journal of the Royal Statistical Society
SERIES C Applied Statistics

Appl. Statist. (2009)
58, *Part* 1, *pp.* 3–24

A Bayesian hierarchical distributed lag model for estimating the time course of risk of hospitalization associated with particulate matter air pollution

Roger D. Peng and Francesca Dominici

Johns Hopkins Bloomberg School of Public Health, Baltimore, USA

and Leah J. Welty

Northwestern University Feinberg School of Medicine, Chicago, USA

[Received June 2007. Revised July 2008]

Summary. Time series studies have provided strong evidence of an association between increased levels of ambient air pollution and increased hospitalizations, typically at a single lag of 0, 1 or 2 days after an air pollution episode. Two important scientific objectives are to understand better how the risk of hospitalization that is associated with a given day's air pollution increase is distributed over multiple days in the future and to estimate the cumulative short-term health effect of an air pollution episode over the same multiday period. We propose a Bayesian hierarchical distributed lag model that integrates information from national health and air pollution databases with prior beliefs of the time course of risk of hospitalization after an air pollution episode. This model is applied to air pollution and health data on 6.3 million enrollees of the US Medicare system living in 94 counties covering the years 1999–2002. We obtain estimates of the distributed lag functions relating fine particulate matter pollution to hospitalizations for both ischaemic heart disease and acute exacerbation of chronic obstructive pulmonary disease, and we use our model to explore regional variation in the health risks across the USA.

Keywords: Air pollution; Cardiovascular disease; Distributed lag model; Environmental epidemiology; Respiratory disease; Time series

1. Introduction

Time series studies of air pollution and health in the USA and around the world have provided consistent evidence of an adverse short-term effect of ambient air pollution levels on mortality and morbidity (Health Effects Institute, 2003; Pope and Dockery, 2006). In particular, multisite studies, which combine information from many locations by using national or regional databases, have produced robust and consistent results demonstrating an adverse health effect that is associated with short-term exposure to particulate matter (PM) and ozone. The National Morbidity, Mortality, and Air Pollution Study in the USA and the 'Air pollution and health: a European approach' study in Europe are prominent examples of such multisite time series studies (Bell *et al.*, 2004; Peng *et al.*, 2005; Katsouyanni *et al.*, 2001; Samoli *et al.*, 2003). More recently, the Medicare Air Pollution Study (MCAPS) showed a strong association between fine PM (PM less than 2.5 μm in aerodynamic diameter) and hospitalization for cardiovascular and respiratory diseases in 204 US counties (Dominici *et al.*, 2006).

Address for correspondence: Roger D. Peng, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, MD 21205, USA.
E-mail: rpeng@jhsph.edu

 0035–9254/09/58003

The majority of previous time series studies of the health effects of PM have generally employed single-lag models that use a fixed exposure lag of l days, assuming that all of the effect of air pollution on health is realized exactly l days in the future. For example, ambient PM levels are often compared with rates of hospitalization on the same day ($l = 0$) or the following day ($l = 1$). Although such an assumption might be plausible for modelling a given individual's response, it is less realistic for describing population level associations since it is unlikely that every member of the population responds to increases in air pollution at the same interval of time. Rather, people will generally respond to an increase in pollution levels at different intervals of time and, when averaged across the population, this response function would appear smooth (Seemungal *et al.*, 2000).

An alternative approach to the single-lag exposure is to use a distributed lag model (DLM) which allows the effect of a single day's increase in air pollution levels to be distributed over multiple days after the increase, thus providing a more informative tool for characterizing the time course of risk of hospitalization. DLMs provide an estimate of the distributed lag function, which describes the change in the relative risk in a multiday period after a given day's increase in air pollution.

DLMs have been used for decades in economics (Almon, 1965; Leamer, 1972; Shiller, 1973) and have been applied more recently in the area of environmental epidemiology. Schwartz (2000) used both unconstrained and constrained (polynomial) distributed lag functions to estimate the effects of PM on daily mortaliy. Zanobetti *et al.* (2000) extended some of this work and developed the generalized additive modelling methodology. Bell *et al.* (2004) and Huang *et al.* (2005) studied the relationship between ozone and daily mortality in the USA and applied both single-lag and constrained DLMs. In general, DLMs will be applicable in areas where the association between an input (or exposure) and a response (or health outcome) can be expected to play out over multiple time points in the future.

DLMs in the area of air pollution and health have primarily been applied to time series data at an individual location such as a county or a city. Typically, a DLM is fitted to the data and the estimated distributed lag function is then smoothed across lags by using a polynomial or non-parametric smoother (e.g. Almon (1965), Corradi (1977) and Zanobetti *et al.* (2000)). Welty *et al.* (2008) proposed a Bayesian model for estimating the distributed lag function in a time series study of a single location. They introduce a prior distribution that constrains the shape of the distributed lag function by allowing effects corresponding to early lags to take a wide range of values whereas effects at more distant lags are constrained to be near zero and correlated with each other. Through extensive simulation studies they showed that their proposed approach is superior (in mean-squared error) to the standard application of penalized splines under several possible shapes of the true distributed lag function. In a problem with potentially many parameters of interest, constraining the distributed lag function in some manner is critical for reducing the size of the model space.

In addition to the distributed lag function, another important target of inference is the cumulative health effect of an increase in air pollution levels over a multiday period after the increase. If the effect of air pollution on health is truly distributed over multiple days, then a relative risk estimate that is obtained by fitting a single-lag model is likely to be biased. However, whether the bias is positive or negative is not clear and either possibility might be considered plausible. For example, Schwartz (2000) found that single-lag models substantially underestimated the effect of PM on daily mortality and similar patterns have been found in other studies (Zanobetti *et al.*, 2002; Goodman *et al.*, 2004; Roberts, 2005). An alternative hypothesis, which is sometimes referred to as the 'harvesting' or 'mortality displacement' hypothesis, claims that air pollution episodes deplete a frail pool of individuals and decrease the number of susceptible people on

future days (Schimmel and Murawski, 1976). Such a phenomenon would lead to a distributed lag function that is negative for certain periods and, when summed over the relevant time period after an air pollution episode, may result in a cumulative effect that is smaller than relative risk estimates obtained from single-lag models (Zeger *et al.*, 1999; Dominici *et al.*, 2002b; Zanobetti *et al.*, 2000).

The National Morbidity, Mortality, and Air Pollution Study, the 'Air pollution and health: a European approach' study and the MCAPS all make clear the substantial advantages of the multisite approach to assessing the short-term health risks of air pollution. Combining information across locations improves the precision of relative risk estimates and allows for the examination of variation in estimates across locations. Hence, there is a need for new methodology to allow the application of DLMs to reap the same benefits.

We introduce a Bayesian hierarchical distributed lag model (BHDLM) for estimating the distributed lag function relating PM air pollution exposure to hospitalizations for cardiovascular and respiratory diseases. We describe a specific prior distribution for constraining the distributed lag function and we propose a hierarchical structure for combining information about the shape of the distributed lag function across multiple locations. We also show a connection between our BHDLM and penalized spline modelling.

We apply our model to a national database of $PM_{2.5}$-measurements and hospitalizations in the USA covering the years 1999–2002. The results include estimates of the national and county-specific distributed lag functions that reflect the contributions of all relevant sources of information as well as their uncertainties. We also obtain estimates of the cumulative effect of $PM_{2.5}$ on hospitalizations in a 2-week period after an increase in levels and compare them with estimates that are obtained from single-lag models. In addition to providing national estimates for the health risks of PM, our model can be used to explore variation in the risks across regions of the country and an assessment of this variation can potentially relate health risks to different sources of PM air pollution. Finally, we provide the R code that was used for fitting this model on the Web site for this paper at `http://www.biostat.jhsph.edu/rr/BHDLM/`.

2. Hierarchical distributed lag model

Given time series data $y_1, y_2, \ldots$ on an outcome such as daily hospitalization counts, and corresponding time series data $x_1, x_2, \ldots$ on an exposure such as ambient air pollution levels, a log-linear Poisson DLM of order L specifies

$$\begin{aligned} y_t &\sim \text{Poisson}(m_t), \\ \log(m_t) &= \sum_{l=0}^{L-1} \mu_l x_{t-l} \end{aligned} \tag{1}$$

for $t \geqslant L-1$. The vector of coefficients

$$\boldsymbol{\mu} = (\mu_0, \mu_1, \ldots, \mu_{L-1}),$$

as a function of the lag number ($l = 0, \ldots, L-1$), is what we call the distributed lag function. This function is sometimes referred to as the impulse–response function because it describes the effect on the outcome series of a single impulse in the exposure series (Chatfield, 1996). For example, if we have an exposure series of the form $x_0 = 1, x_1 = 0, x_2 = 0, \ldots$, i.e. a spike at $t = 0$, then the log-relative risk (over L days) that is associated with that spike is $\xi = \Sigma_{l=0}^{L-1} \mu_l$. We define $100\{\exp(10\xi) - 1\}$ to be the cumulative percentage increase in hospitalizations over an L-day period associated with an increase in pollution of $10\,\mu\text{g m}^{-3}$ (a standard increment for reporting particle air pollution relative risks).

2.1. Single-county model

Our approach begins with a model for air pollution and hospitalization data in a single location such as a county. This model relates day-to-day changes in air pollution levels to day-to-day changes in rates of hospitalization for a given county, controlling for other time varying factors that might confound the relationship of interest.

Let the vector $\mathbf{y} = (y_1, y_2, \ldots, y_T)$ represent the daily time series of hospitalization counts for a county and let the vector $\mathbf{d}$ be the daily time series of the numbers of people at risk. The matrix X represents the exposure of interest and includes the corresponding time series of air pollution levels and lagged versions of that series. X is of dimension $T \times L$, where L is the order of the DLM. In our set-up, the first column of X is the original air pollution time series (lag 0), the second column is the original series lagged by 1 day (lag 1), etc. We also observe p other time varying covariates which are combined in a $T \times p$ matrix Z. Then our single-county log-linear Poisson model is of the form

$$\begin{aligned} \mathbf{y}|X, Z &\sim \text{Poisson}\{m(\boldsymbol{\mu}, \boldsymbol{\beta})\}, \\ \log\{m(\boldsymbol{\mu}, \boldsymbol{\beta})\} &= X\boldsymbol{\mu} + Z\boldsymbol{\beta} + \log(\mathbf{d}) \end{aligned} \tag{2}$$

where the length L vector of parameters $\boldsymbol{\mu}$ is the distributed lag function and parameters in $\boldsymbol{\beta}$ are nuisance parameters.

In the county-specific model (2), we incorporate into Z certain time varying factors that might confound the relationship between air pollution and hospitalization (Kelsall *et al.*, 1997; Dominici *et al.*, 2002a). In particular, we include smooth functions of average daily temperature, dewpoint temperature and indicators for the day of the week. We also include a smooth function of time to adjust for seasonal variation that is common to both the air pollution and the hospitalization time series. This smooth function of time is modelled by using natural splines and the natural spline basis is included in Z. Further details regarding the approaches to confounding adjustment in time series studies of air pollution and health can be found in Peng *et al.* (2006) and Welty and Zeger (2005).

2.2. Constraining the distributed lag function

The rationale behind our approach to constraining the distributed lag function $\boldsymbol{\mu}$ is that the effects of air pollution at early lags are not well understood because of our lack of knowledge about biological mechanisms and the time course of the disease process within the population. In addition, competing hypotheses that were mentioned previously about the shape of the distributed lag function suggest that fewer constraints should be placed at early lags.

At longer lags there are both substantive and methodological justifications for assuming that the effects of air pollution on the outcome should approach zero smoothly. First, there is little evidence, epidemiological or toxicological, that short-term increases in air pollution have effects that extend beyond a few days or a week. For example, toxicological experiments conducted in rats and humans indicate that symptoms of inflammation generally resolve within a week (Clarke *et al.*, 1999; Lay *et al.*, 1999). Even in such an extreme scenario as the London fog of 1952, where pollution reached levels that were 100 times today's levels, the bulk of excess mortality occurred within 2 weeks of the initial event (Bell and Davis, 2001).

The second justification for assuming that the effects of air pollution approach zero as the lag increases is methodological and concerns the potential confounding by season. Although there may indeed be longer-term effects of short-term increases in air pollution, say of the order of many weeks or months, it is difficult to disentangle such effects from the much more powerful effects of seasonal factors such as influenza epidemics and other unmeasured phenomena. As mentioned in Section 2.1, we typically include a smooth function of time in the county-spe-

cific model to remove such seasonal variation from the data. Once variation in $PM_{2.5}$ at longer timescales has been removed, we cannot hope to estimate associations that extend into those timescales. Hence, the choice of how long a lag to examine is closely connected to the adjustment for potential confounding by seasonal factors. Our choice of which timescales of variation to retain and which to remove is based on extensive previous work and sensitivity analysis in this area (Peng *et al.*, 2006; Dominici *et al.*, 2003; Zeger *et al.*, 1999).

For a single county, we can constrain the distributed lag function $\boldsymbol{\mu}$ by specifying a prior distribution

$$\boldsymbol{\mu}|\boldsymbol{\gamma} \sim \mathcal{N}\{\mathbf{0}, \sigma_\gamma^2\, \Omega(\boldsymbol{\gamma})\} \tag{3}$$

where the vector $\boldsymbol{\mu} = (\mu_0, \mu_1, \ldots, \mu_{L-1})$ contains the values of the distributed lag function at each lag l and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$ controls the nature of the constraint on the distributed lag function.

The covariance matrix Ω is parameterized by the vector $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$ where γ_1 controls the rate at which the variance of the distributed lag function coefficients taper to 0 and γ_2 controls the rate at which neighbouring coefficients become more correlated. Specifically, we assume that the variance of μ_l tapers to 0 exponentially as a function of l, so that $\text{var}(\mu_l) = \sigma_\gamma^2 \exp(-\gamma_1 l)$ for $l = 0, 1, 2, \ldots, L-1$. We further assume that the covariance of neighbouring coefficients at lags l_1 and l_2 (for $l_1 \neq l_2$) is

$$\text{cov}(\mu_{l_1}, \mu_{l_2}) = \frac{\sigma_\gamma^2\{1-\exp(-\gamma_2 l_1)\}\{1-\exp(-\gamma_2 l_2)\}\ \exp\{-\gamma_1(l_1+l_2)/2\}}{\sqrt{([\{1-\exp(-\gamma_2 l_1)\}^2+\exp(-2\gamma_2 l_1)][\{1-\exp(\gamma_2 l_2)\}^2+\exp(-2\gamma_2 l_2)])}}$$

so that neighbouring coefficients at large lags have a correlation that is close to 1. Greater detail about the construction of this covariance matrix can be found in Welty *et al.* (2008). The parameter σ_γ^2 is the prior variance of μ_0, the first distributed lag coefficient.

Samples from the prior distribution in distribution (3) for various values of γ_1 and γ_2 are shown in Fig. 1. The possibilities range from completely uncorrelated lag effects in Fig. 1(a) ($\gamma_1 = \gamma_2 = 0$) to highly constrained lag effects in Fig. 1(i) ($\gamma_1 = \gamma_2 = 1$). Given enough prior information one could feasibly choose fixed values for γ_1 and γ_2 or, lacking such information, place a hyperprior distribution on these parameters, as we do (see Section 2.3.1).

2.3. Combining information across counties

The approach that was described in Section 2.2 is limited to estimating the distributed lag function for a single county. In this section we extend the model to handle situations when time series data are available from multiple counties.

At the county level, we use the log-linear Poisson model for the county-specific hospital admissions rates and air pollution data. For each county $c = 1, \ldots, n$, we have

$$\begin{aligned} \mathbf{y}_c|X_c, Z_c &\sim \text{Poisson}\{m_c(\boldsymbol{\theta}_c, \boldsymbol{\beta}_c)\}, \\ \log\{m_c(\boldsymbol{\theta}_c, \boldsymbol{\beta}_c)\} &= X_c\boldsymbol{\theta}_c + Z_c\boldsymbol{\beta}_c + \log(\mathbf{d}_c). \end{aligned} \tag{4}$$

A direct extension of the single-county model would assume a common association between air pollution and hospital admissions across counties. Rather than assume a common association, we allow each county to have its own parameter $\boldsymbol{\theta}_c$ to allow for some heterogeneity between counties.

We assume that each of the county-specific distributed lag functions $\boldsymbol{\theta}_c$ is normally distributed around a 'national average' distributed lag function $\boldsymbol{\mu}$,

$$\boldsymbol{\theta}_c|\boldsymbol{\mu}, \boldsymbol{\eta}, \sigma_\eta^2 \sim \mathcal{N}\{\boldsymbol{\mu}, \sigma_\eta^2\, \Omega(\boldsymbol{\eta})\} \tag{5}$$

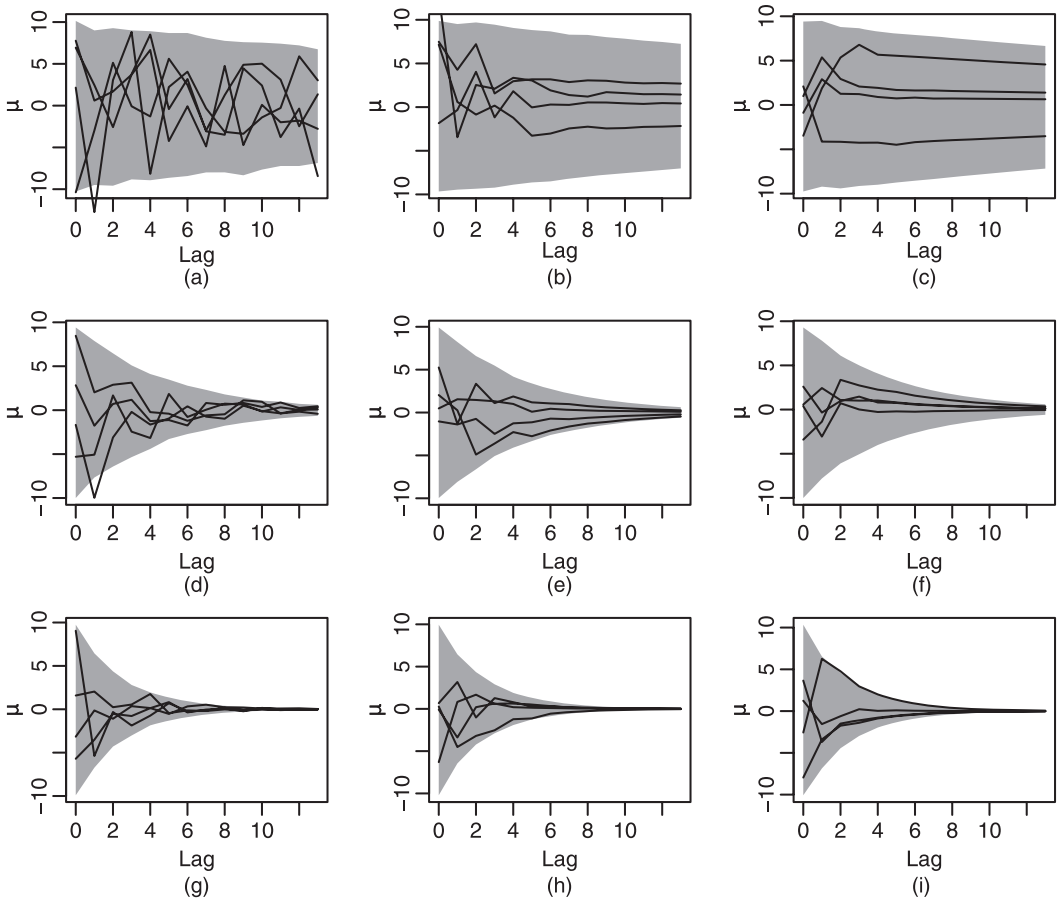

Fig. 1. Samples from the prior distribution for the constrained distributed lag function (grey regions indicate pointwise 95% intervals; a model with $\gamma_1 = \gamma_2 = 0$ has uncorrelated lag effects whereas a model with $\gamma_1 = \gamma_2 = 1$ is highly constrained): (a) $\gamma_1 = 0$ and $\gamma_2 = 0$; (b) $\gamma_1 = 0$ and $\gamma_2 = 0.5$; (c) $\gamma_1 = 0$ and $\gamma_2 = 1$; (d) $\gamma_1 = 0.5$ and $\gamma_2 = 0$; (e) $\gamma_1 = 0.5$ and $\gamma_2 = 0.5$; (f) $\gamma_1 = 0.5$ and $\gamma_2 = 1$; (g) $\gamma_1 = 1$ and $\gamma_2 = 0$; (h) $\gamma_1 = 1$ and $\gamma_2 = 0.5$; (i) $\gamma_1 = 1$ and $\gamma_2 = 1$

where the matrix Ω is constructed in the same manner as described in Section 2.2 except now parameterized by the vector $\boldsymbol{\eta} = (\eta_1, \eta_2)$. The matrix $\Omega(\boldsymbol{\eta})$ describes the unexplained variation or heterogeneity across counties of the county-specific distributed lag functions $\boldsymbol{\theta}_c$.

The interpretation of $\Omega(\boldsymbol{\eta})$ here is that we assume *a priori* that there will be more variation across counties in the coefficients corresponding to early lags and less variation in the coefficients corresponding to longer lags. Large values of $\sigma_{\boldsymbol{\eta}}^2$ allow for inferences about early lag values to be weighted in favour of the county-specific data; the amount of shrinkage towards the national average increases with lag. Smaller values of $\sigma_{\boldsymbol{\eta}}^2$ favour more shrinkage towards the national average at early lags. Here, we specifically take advantage of the multisite context by exploring the variation in the shapes of the county-specific distributed lag functions. Such unexplained variation may exist, for example, because of varying composition of PM across the country or because of varying susceptibilities of the populations.

We assume that the prior distribution for the national-average-distributed lag function $\boldsymbol{\mu}$ is

$$\boldsymbol{\mu} | \boldsymbol{\gamma}, \sigma_{\gamma}^2 \sim \mathcal{N}\{\mathbf{0}, \sigma_{\gamma}^2 \, \Omega(\boldsymbol{\gamma})\}, \tag{6}$$

where Ω has the same form as in distribution (5) but is parameterized by the vector $\gamma = (\gamma_1, \gamma_2)$ and σ_γ^2 is the prior variance of μ_0, the national average effect at lag 0. This formulation is analogous to the single-county model in Section 2.2.

2.3.1. Hyperprior specification

To complete the model specification, the parameters η and γ are each assumed to have uniform hyperprior distributions over a fixed range. The ranges of the η- and γ-parameters allow for an unconstrained distributed lag function as well as some heavily constrained and smooth distributed lag functions (details of the ranges are given in Appendix A). Exploratory analyses indicated that there was little information in the data to estimate jointly σ_η and σ_γ as well as η and γ (see for example Schmidt *et al.* (2007)). Therefore, we set σ_η and σ_γ to be approximately 10 times the square root of the variance of the maximum likelihood estimate of μ_0. These values of σ_η and σ_γ ensure that, even for highly constrained models (i.e. large values of η or γ), the prior has little influence over the coefficients corresponding to the early lags. Sensitivity analysis indicates that the relevant posterior distributions are not substantially affected as long as the values of σ_η and σ_γ are not too small (Welty *et al.*, 2008).

We implement a Gibbs sampler to obtain samples from the posterior distributions of the unknown parameters μ, γ, η and θ_c for $c = 1, \ldots, n$. Full details of the sampling procedures can be found in Appendix A.

2.4. Two-stage distributed lag models

One approximation to our BHDLM that was described in Section 2.3 is the following two-stage DLM. At the first stage, a distributed lag function is estimated independently for each location by using log-linear Poisson regressions. This first-stage distributed lag function can be either constrained or unconstrained. Given the estimates of the distributed lag function for each county $\hat{\theta}_c$ and their covariances $\hat{\Sigma}_c$ obtained by using county-specific regressions, we use a normal approximation to the Poisson likelihood and we assume that

$$\hat{\theta}_c | \theta_c \sim \mathcal{N}(\theta_c, \hat{\Sigma}_c).$$

Given the normal approximation to the likelihood, we further assume that

$$\theta_c | \mu \sim \mathcal{N}(\mu, \Psi),$$

where Ψ is an unconstrained $L \times L$ covariance matrix. Estimates of μ that are obtained by using the two-stage approach are analogous to those obtained by using the BHDLM that was described in Section 2.3. The two-stage approach has been used in numerous previous studies and is attractive for its computational simplicity.

For comparison with our model, we implemented a specific case of the two-stage approach that was described above. Within each county c, we assume a constrained distributed lag function

$$\log(\mathbb{E}[Y_t^c]) = \theta_0^c \bar{x}_{t,0-2}^c + \theta_1^c \bar{x}_{t,3-6}^c + \theta_2^c \bar{x}_{t,7-13}^c + \text{other predictors}$$

where

$$\bar{x}_{t,0-2}^c = \frac{1}{3} \sum_{l=0}^{2} x_{t-l}^c,$$

$$\bar{x}_{t,3-6}^c = \frac{1}{4} \sum_{l=3}^{6} x_{t-l}^c$$

and

$$\bar{x}^c_{t,7-13} = \frac{1}{7}\sum_{l=7}^{13} x^c_{t-l}$$

are simply running means of the specified lengths. This model is constrained at the county level in the sense that the effects of pollution at lags 0–2, lags 3–6 and lags 7–13 are restricted to be constant respectively, so that the distributed lag function resembles a step function. We have chosen this model for comparison largely because of the simplicity and computational efficiency with which we can estimate the parameters. In addition, this type of model has been used previously in the literature (Bell *et al.*, 2004; Welty and Zeger, 2005).

Posterior samples for the parameters $\boldsymbol{\mu}$ and Ψ for this model are obtained by using the two-level normal independent sampling estimation software of Everson and Morris (2000). Although the county-specific DLM that is specified here is constrained to be a step function, the second-level covariance matrix Ψ here is completely unconstrained. This approach differs from our model, where the county-specific function in expression (4) can vary freely but the second-level covariance matrix $\Omega(\boldsymbol{\eta})$ is constrained.

2.5. Connection to penalized splines

Our BHDLM can be reformulated as a penalized spline model where the prior distributions in expressions (5) and (6) induce a special type of penalty for constraining the county-specific distributed lag functions and combining information across counties. To show this connection, we shall use the normal approximation of our model to allow the computations to be written in closed form.

Let $\hat{\boldsymbol{\theta}}_c$ and $\hat{\Sigma}_c$ be the maximum likelihood estimates of the distributed lag coefficients and the corresponding covariance matrix for county c. We shall assume a normal distribution for $\hat{\boldsymbol{\theta}}_c$ so that the estimated distributed lag function $\hat{\boldsymbol{\theta}}_c$ can be modelled as a linear combination of basis functions, $\hat{\boldsymbol{\theta}}_c = U\boldsymbol{\alpha}_c + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \hat{\Sigma}_c)$. U is an $L \times k$ basis matrix and $\boldsymbol{\alpha}_c$ is a k-vector of coefficients. The penalized spline solution solves the following optimization problem:

$$\min_{\boldsymbol{\alpha}_c}\{(\hat{\boldsymbol{\theta}}_c - U\boldsymbol{\alpha}_c)'\hat{\Sigma}_c^{-1}(\hat{\boldsymbol{\theta}}_c - U\boldsymbol{\alpha}_c) + \boldsymbol{\alpha}_c' D^{-1}\boldsymbol{\alpha}_c\},$$

where D^{-1} is a penalty matrix which we assume incorporates a scalar penalty parameter. Since the penalty term $\boldsymbol{\alpha}_c' D^{-1}\boldsymbol{\alpha}_c$ is proportional to the minus log-density of a normal distribution, we can rewrite the problem as

$$\begin{aligned} \hat{\boldsymbol{\theta}}_c | U\boldsymbol{\alpha}_c &\sim \mathcal{N}(U\boldsymbol{\alpha}_c, \hat{\Sigma}_c), \\ \boldsymbol{\alpha}_c &\sim \mathcal{N}(\mathbf{0}, D) \end{aligned} \tag{7}$$

where the solution is the posterior mode of $\boldsymbol{\alpha}_c$ under the normal prior in distribution (7).

Given expressions (5) and (6), we can write the marginal distribution of $\boldsymbol{\theta}_c$ as

$$\boldsymbol{\theta}_c \sim \mathcal{N}\{\mathbf{0}, \Omega(\boldsymbol{\eta}) + \Omega(\boldsymbol{\gamma})\} \tag{8}$$

where we have absorbed $\sigma^2_{\boldsymbol{\eta}}$ and $\sigma^2_{\boldsymbol{\gamma}}$ into $\Omega(\boldsymbol{\eta})$ and $\Omega(\boldsymbol{\gamma})$ respectively, to reduce the clutter. The distribution for $\boldsymbol{\alpha}_c$ in expression (7) implies that $U\boldsymbol{\alpha}_c \sim \mathcal{N}(\mathbf{0}, UDU')$. On the basis of our previous notation, $\boldsymbol{\theta}_c = U\boldsymbol{\alpha}_c$, so the (inverse) penalty matrix D must satisfy $UDU' = \Omega(\boldsymbol{\eta}) + \Omega(\boldsymbol{\gamma})$, which has the solution

$$D_{\boldsymbol{\eta},\boldsymbol{\gamma}} = (U'U)^{-1}U'\{\Omega(\boldsymbol{\eta}) + \Omega(\boldsymbol{\gamma})\}U(U'U)^{-1}. \tag{9}$$

Now we have shown that our prior distribution on $\boldsymbol{\theta}_c$ can be translated into a penalty matrix

for spline coefficients $\boldsymbol{\alpha}_c$ in a penalized spline problem. Given values of $\boldsymbol{\eta}$ and γ and using the penalty matrix in equation (9), we can calculate the penalized spline coefficient estimates as $\hat{\boldsymbol{\alpha}}_c = D_{\boldsymbol{\eta},\gamma} U'(U D_{\boldsymbol{\eta},\gamma} U' + \hat{\Sigma}_c)^{-1} \hat{\boldsymbol{\theta}}_c$ and the smoothed county-specific distributed lag function for county c is $U\hat{\boldsymbol{\alpha}}_c$.

Similar calculations demonstrate how penalized splines can be used to combine the county-specific distributed lag functions $\boldsymbol{\theta}_c$ across counties. Analogously to expression (7), we can write the second level of our hierarchical model as

$$\begin{aligned} \boldsymbol{\theta}_c | W\boldsymbol{\delta} &\sim \mathcal{N}\{W\boldsymbol{\delta}, \Omega(\boldsymbol{\eta})\} \qquad (c = 1, \ldots, n), \\ \boldsymbol{\delta} &\sim \mathcal{N}(\mathbf{0}, H) \end{aligned} \tag{10}$$

where W is a spline basis matrix, $\boldsymbol{\delta}$ is a vector of coefficients and H is a penalty matrix in the penalized spline problem. The distribution in expression (6) and the prior for $\boldsymbol{\delta}$ in expression (10) imply that we need to find a matrix H such that $WHW' = \Omega(\gamma)$. The solution for H has the form

$$H_\gamma = (W'W)^{-1} W' \, \Omega(\gamma) W (W'W)^{-1}$$

and we can subsequently solve for

$$\hat{\boldsymbol{\delta}} = H_\gamma W' \{W H_\gamma W' + \Omega(\boldsymbol{\eta})\}^{-1} \bar{\boldsymbol{\theta}},$$

where $\bar{\boldsymbol{\theta}} = (1/n) \Sigma_{c=1}^{n} \boldsymbol{\theta}_c$.

We can see that, if we replace the basis matrices U and W with the $L \times L$ identity matrix, then we revert to our original formulation and obtain the same answers as our original Bayesian hierarchical model. Our model places a prior directly on the distributed lag function, whereas the penalized spline approach places a prior on the corresponding spline coefficients. In this application it seems more natural to assume a prior distribution for the distributed lag function directly because prior information is available in that domain.

3. Data

We apply our methods to national databases of hospitalization and ambient $PM_{2.5}$-measurements. The hospitalization data consist of daily counts of hospital admissions for the years 1999–2002 constructed from the national claims history files of the US Medicare system, which contain the billing claims of all Medicare enrollees. Medicare enrollees make up almost the entire US population over 65 years of age, or approximately 48 million people. Each billing claim that was obtained from the national claims history files contains the date of service, treatment, disease classification (via international classification of diseases, ninth revision, codes), age, gender, self-reported race and place of residence (five-digit zip code and county). The daily counts for a given county were computed by summing the total number of hospitalizations with a primary diagnosis for a specific disease. For computing hospitalization rates, a corresponding time series of the numbers of individuals at risk in each county for each day was constructed.

The $PM_{2.5}$-data were obtained from the US Environmental Protection Agency's air quality system database, which makes available data from a national network of monitors. Before 1999, the Environmental Protection Agency collected data on PM that was less than 10 μm in diameter (PM_{10}), generally on a 1-in-6 day basis (i.e., for every 6 days, one measurement of PM_{10} is made). Such a data collection pattern ruled out the use of DLMs, which require data to be collected on consecutive days (although daily PM_{10}-data were collected for about 15 counties).

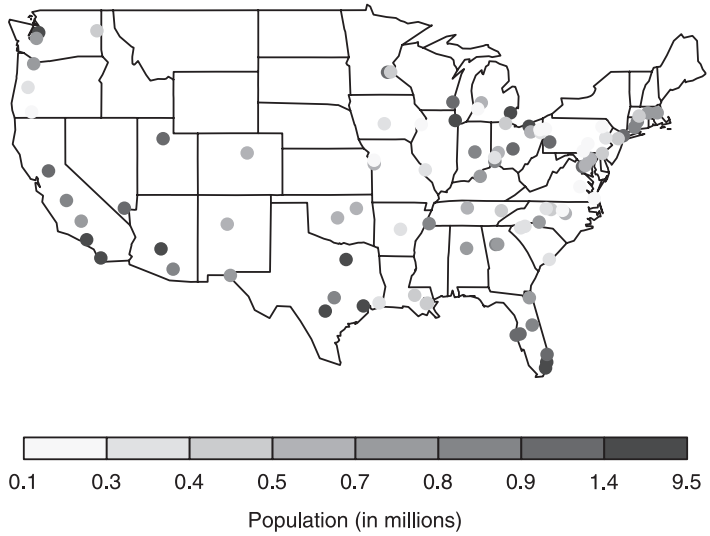

Fig. 2. Locations of 94 US counties which have daily data for PM less than 2.5 μm in diameter for 1999–2002

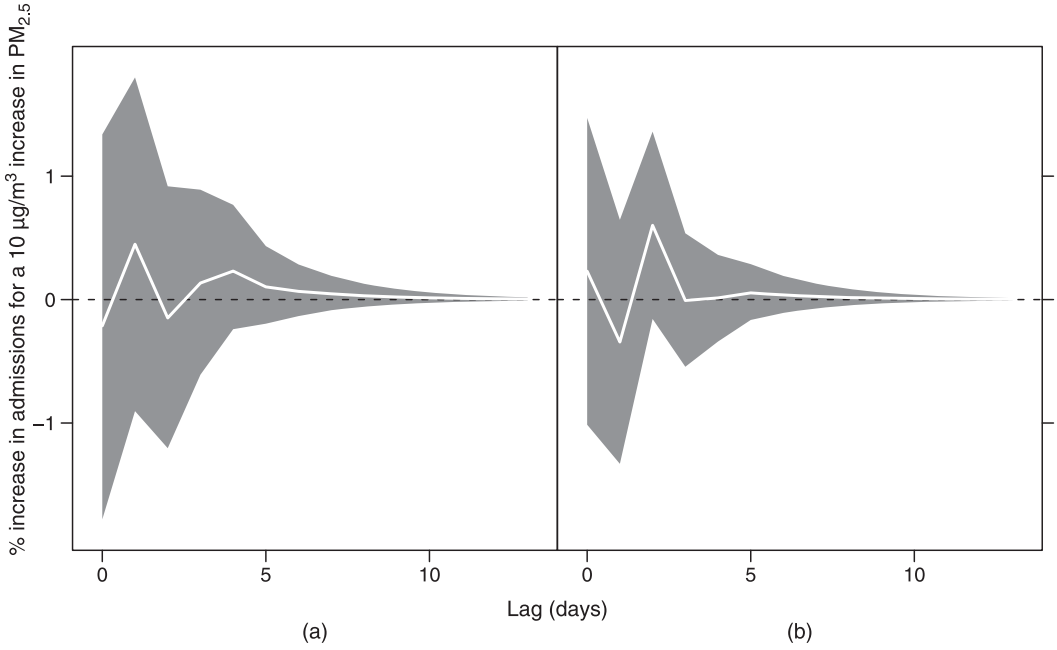

Fig. 3. National average distributed lag functions for (a) COPDAE and (b) ischaemic heart disease from the BHDLM applied to 94 US counties, 1999–2002: each plot shows the posterior mean (white curve) and pointwise 95% posterior intervals (grey regions) for each lag coefficient

Beginning in 1999, the Environmental Protection Agency began to collect $PM_{2.5}$-data, generally on a 1-in-3 day basis, although there are over 100 counties where measurements are taken every day. With the emergence of the new $PM_{2.5}$ monitoring network, we can fit DLMs to data from more counties than previously possible. Some counties contained more than one $PM_{2.5}$-monitor, in which case we took a 10% trimmed mean of the daily values across monitors. In cases where there were fewer than 10 monitor readings, we dropped the lowest and highest values and

averaged the remainder. Finally, temperature and dewpoint temperature data were assembled from the National Climatic Data Center on the Earth-Info compact disc database and linked by county with the pollution and hospitalization data.

This analysis was necessarily restricted to the counties for which daily data on $PM_{2.5}$ were available. The counties included were further constrained to have a population over 200 000 and $PM_{2.5}$-data spanning at least one full year. The resulting study population resided in 94 counties and consisted of 6.3 million Medicare enrollees living on average 6 miles from a $PM_{2.5}$-monitor. The locations and populations of the 94 counties are shown in Fig. 2.

For all the counties that were used in this analysis, there were occasional missing $PM_{2.5}$-values. With the exception of the year 1999, when monitors in some counties were just beginning to come into service, the missingness tended to be sporadic and seemingly at random. Rather than treat the missing $PM_{2.5}$-values specially or implement an imputation scheme, we chose simply to drop missing observations and to analyse only the days for which observations were available. One issue that arises by taking this approach is that, when fitting a DLM of order L, a single missing value in the exposure series propagates to create L missing observations in the health effects model. Fortunately, the number of missing $PM_{2.5}$-values was sufficiently small that this propagation did not cause a serious problem. There were few, if any, missing values in the hospitalization and meteorological data.

4. Results

We applied the BHDLM to the 94 counties with Medicare, air pollution and weather data that were described in Section 3. The data for each county spanned $T = 1461$ days (the four years from 1999 to 2002) and we chose to examine two specific causes of hospitalization: chronic obstructive pulmonary disease with acute exacerbation (COPDAE) and ischaemic heart disease. These outcomes were chosen because they represent common respiratory and cardiovascular diseases and have been shown in previous studies to be strongly associated with $PM_{2.5}$-exposure. For the distributed lag function, we chose to fit a model with a maximum lag of 2 weeks, so that

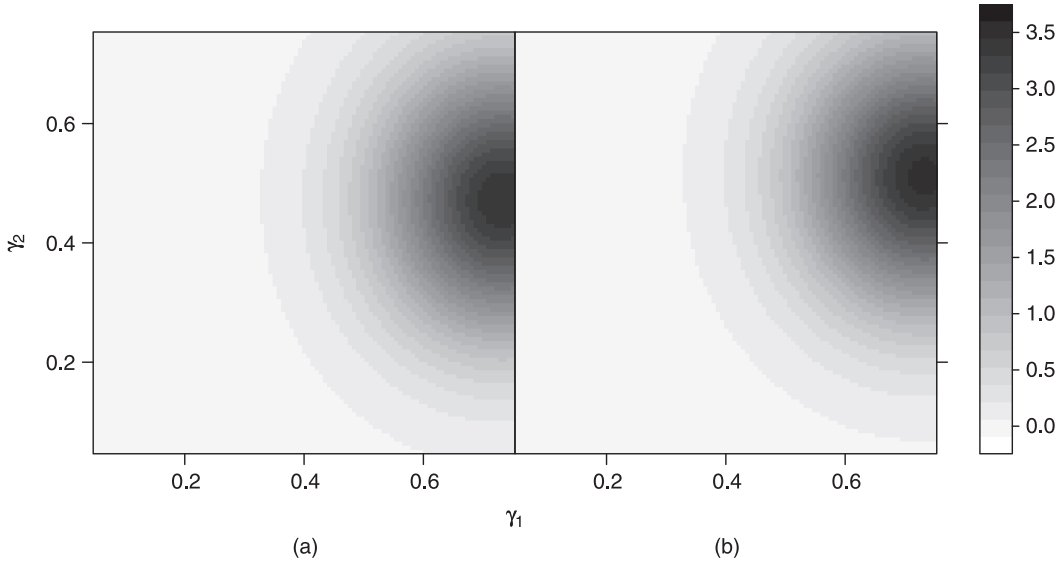

Fig. 4. Joint marginal posterior distributions for γ_1 and γ_2 for both (a) COPDAE and (b) ischaemic heart disease

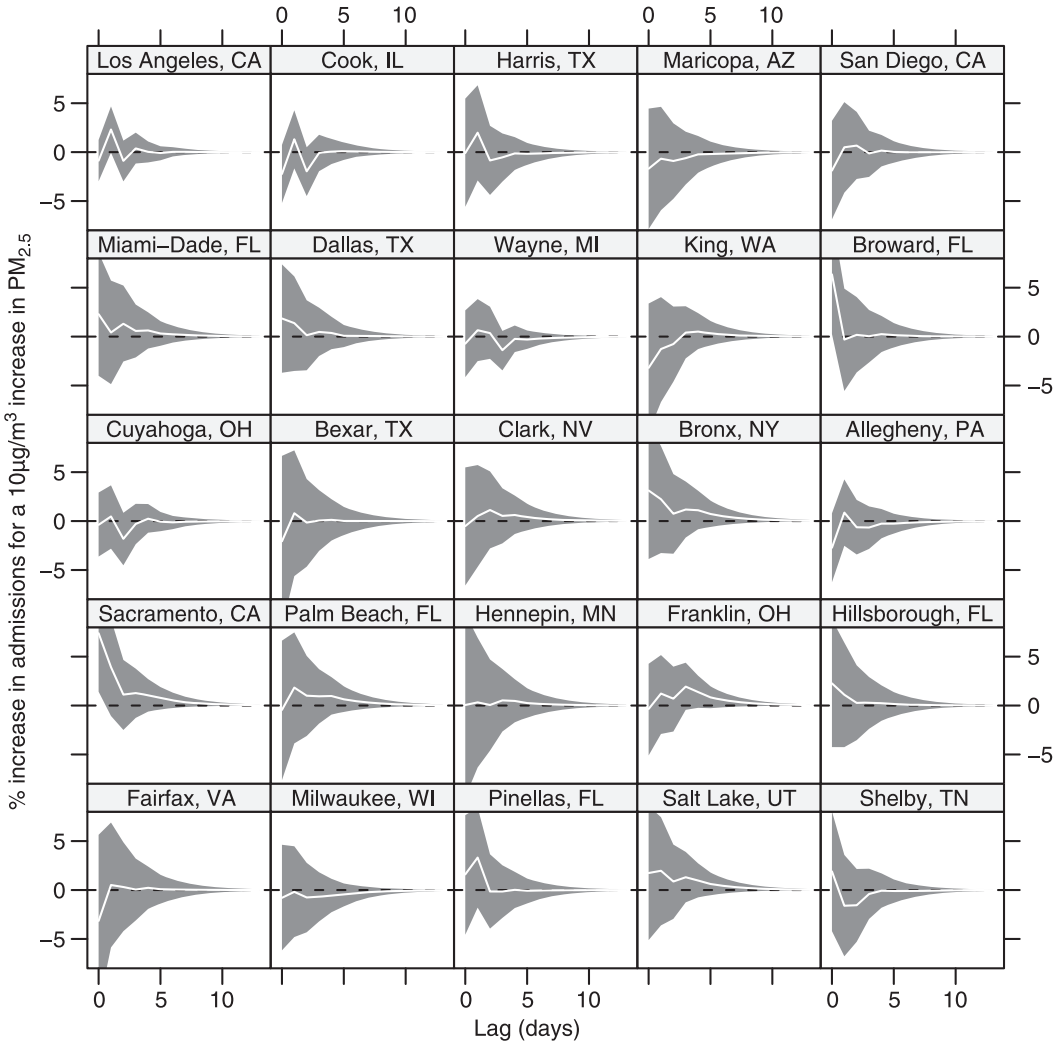

Fig. 5. County-specific Bayesian distributed lag functions (with pointwise 95% posterior intervals) showing the effect of $PM_{2.5}$ on hospitalization for COPDAE: only the largest 25 counties (by population) are shown here, with the largest county (Los Angeles, California) in the top left-hand corner

$L = 14$ in model (1). As a comparison we also applied the step function DLM (step DLM) that was described in Section 2.4 where the model is fitted by using the two-stage approach.

The national average distributed lag functions that were estimated by the BHDLM for COPDAE and ischaemic heart disease are shown in Fig. 3. Each of these plots shows the posterior mean for $\boldsymbol{\mu}$ plotted as a function of lag for each outcome as well as pointwise 95% posterior intervals for each lag coefficient. At each lag the plotted coefficient can be interpreted as the percentage increase in hospitalization for an increase in $PM_{2.5}$ of 10 μg m^{-3}. For COPDAE, Fig. 3 suggests that $PM_{2.5}$ is associated with two 'waves' of admissions, with the first arriving 1 day after the increase and the second arriving a few days later. For ischaemic heart disease, there is an increase in admissions of about 0.23% on the same day, followed by a decrease in admissions of approximately 0.34% on the following day. At lag 2, the relative risk jumps to an

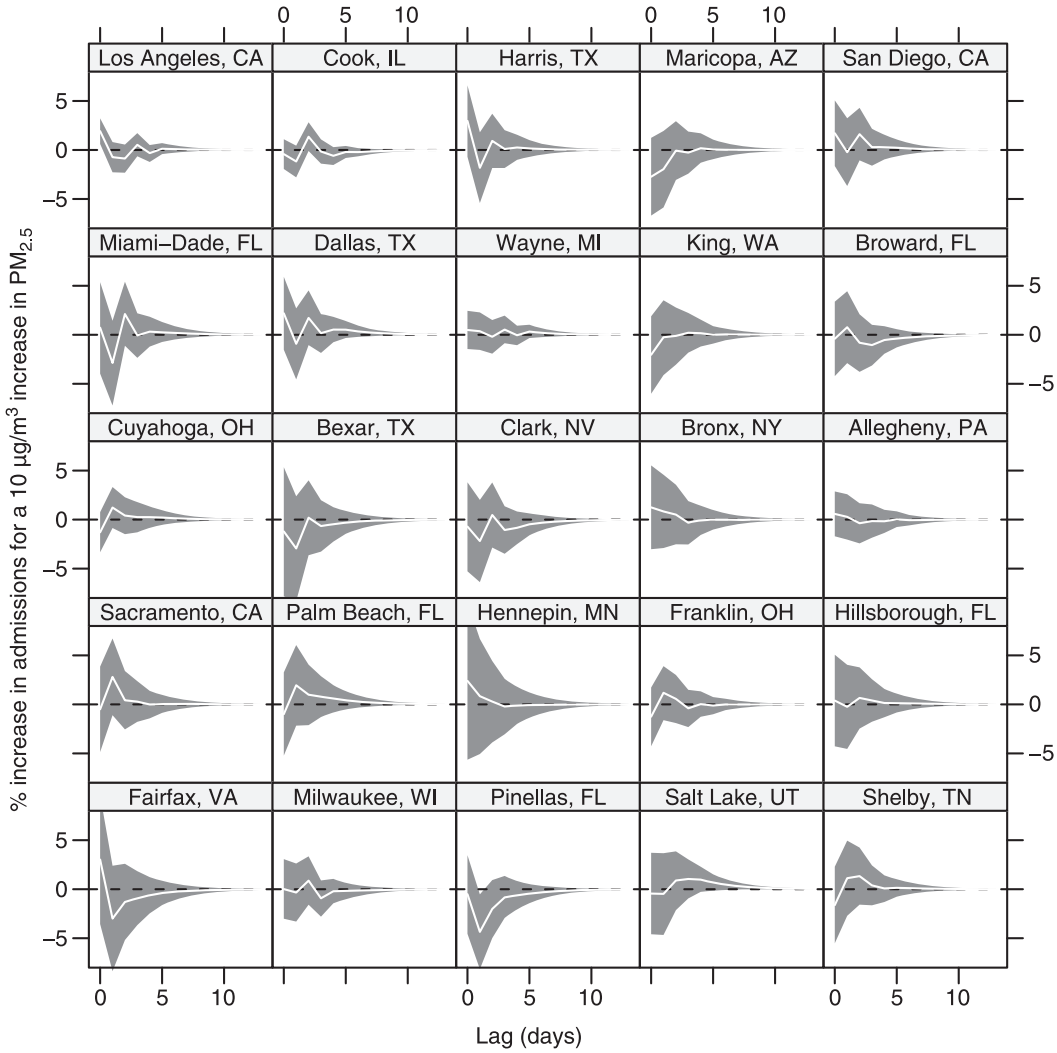

Fig. 6. County-specific Bayesian distributed lag functions (with pointwise 95% posterior intervals) showing the effect of $PM_{2.5}$ on hospitalization for ischaemic heart disease: only the largest 25 counties (by population) are shown here, with the largest county (Los Angeles, California) in the top left-hand corner

increase in admissions of 0.6%, beyond which the distributed lag function for ischaemic heart disease is essentially zero.

The joint marginal posterior distributions for γ_1 and γ_2, which control the tapering and smoothness of $\boldsymbol{\mu}$, are shown in Fig. 4 for both COPDAE and ischaemic heart disease. Large values of γ_1 indicate a strong tapering of the lag coefficients towards 0 whereas large values of γ_2 indicate a very smooth distributed lag function. The data for both outcomes prefer a large value for γ_1, indicating strong variance tapering, but for ischaemic heart disease the marginal distribution for γ_2 is shifted somewhat higher than that of COPDAE.

The county-specific Bayesian distributed lag functions for COPDAE and ischaemic heart disease are shown in Figs 5 and 6 respectively. Each figure shows the posterior mean and pointwise 95% posterior intervals of $\boldsymbol{\theta}_c$ for the largest 25 counties in the study. For COPDAE, the estim-

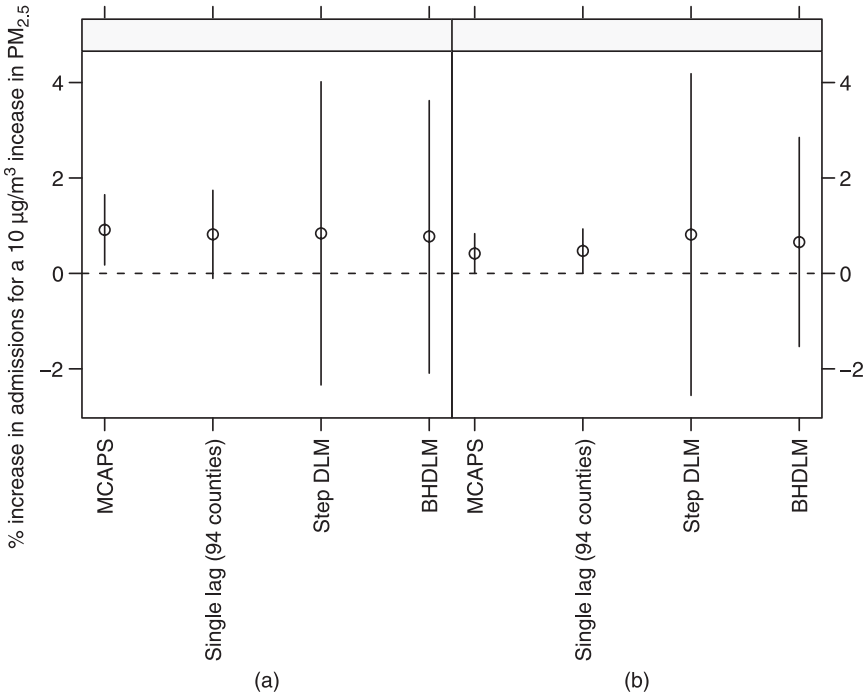

Fig. 7. Estimates and 95% posterior intervals for the cumulative effect of $PM_{2.5}$ for (a) COPDAE and (b) ischaemic heart disease: estimates for 'MCAPS' and 'single lag (94 counties)' come from single-lag models applied to the original MCAPS and to the 94 counties that were used in this study respectively (lag 0 for COPDAE and lag 2 for ischaemic heart disease); the 'step DLM' estimates come from a 14-day DLM using a step function (county-specific estimates are pooled by using the two-stage approach); the 'BHDLM' estimates come from applying the BHDLM using a 14-day distributed lag

ated county-specific distributed lag functions are a mix of shapes including large immediate effects (Sacramento, California, and Broward, Florida), somewhat smaller delayed effects (Los Angeles, California, Franklin, Ohio, and Pinellas, Florida), and more moderate effects spread out over a longer period of time (Bronx, New York, Palm Beach, Florida, and Salt Lake, Utah). Ischaemic heart disease appears to exhibit somewhat less heterogeneity in the shapes of the distributed lag functions with most of the effects occurring at lags 0–2. In Fairfax, Virginia, and Pinellas, Florida, counties there appears to be some evidence of mortality displacement.

Fig. 7 shows the posterior mean and 95% posterior intervals of the association between $PM_{2.5}$ and both outcomes. The estimates were obtained from two single-lag models and two DLMs. In the single-lag models we report the estimate for a given lag and for the DLMs we report the cumulative effect over 14 days. For each outcome we plot

(a) the estimate that was originally reported in the MCAPS for a single-lag model applied to 204 US counties ('MCAPS'),
(b) the estimate that was obtained from a single-lag model ('single lag') applied to the 94 counties that were used in this study (for the exposure lag we chose lag 0 for COPDAE and lag 2 for ischaemic heart disease, the same lags as used in the MCAPS),
(c) the estimate that was obtained from using the step function model via the two-stage approach ('step DLM') and
(d) the estimate that was obtained from our BHDLM ('BHDLM').

For the COPDAE outcome, the MCAPS point estimate was reported as 0.91 with a 95% posterior interval of (0.18, 1.64) and the posterior mean from the BHDLM model was 0.77 (−2.09, 3.62). We can see from the difference in posterior intervals from the 'MCAPS' and the 'single-lag' estimates that the loss of 110 counties in this study results in only a small loss of efficiency in the estimate of the single-lag effect.

In the DLM, the increased number of parameters introduced (even in the three-parameter step DLM model) results in a substantial increase in the variance of the cumulative effect estimate. For the ischaemic heart disease outcome, the estimate from the BHDLM is 0.66 (−1.53, 2.85) compared with the MCAPS estimate of 0.44 (0.02, 0.86). This higher effect was also captured by the step function DLM but the estimate from the BHDLM appears to exhibit less variance in its estimate.

4.1. Regional variation

One significant advantage of our analysis is that it provides the opportunity to examine variation in the county-specific distributed lag functions across locations and regions. Variation in the composition of $PM_{2.5}$ may correspond to variation in the estimated distributed lag functions if the concentrations of the most toxic constituents of the PM mixture vary across locations (Bell *et al.*, 2007a; Lippmann *et al.*, 2006). Variation in the estimated relative risks may also indicate regional variation in the susceptibilities in the underlying populations to exposure to $PM_{2.5}$.

We compared the estimated county-specific distributed lag functions for 53 counties in the northern region of the USA (defined as having latitude above 36.5°) with 41 counties in the southern region of the USA to see whether there were any systematic differences. Given the N posterior samples of the county-specific distributed lag functions $\boldsymbol{\theta}_c$, we calculated regional averages

$$\boldsymbol{\theta}_R^{(i)} = \frac{1}{n_R} \sum_{c \in \mathcal{I}_R} \boldsymbol{\theta}_c^{(i)}$$

for each of the posterior samples i, where R indicates the region (north or south) and $\mathcal{I}_R$ is the index set for the counties in region R. The posterior mean $\bar{\boldsymbol{\theta}}_R$ was then computed by averaging $\boldsymbol{\theta}_R^{(i)}$ over the posterior samples.

Examination of the cumulative effects for the north and south show a clear regional differ-

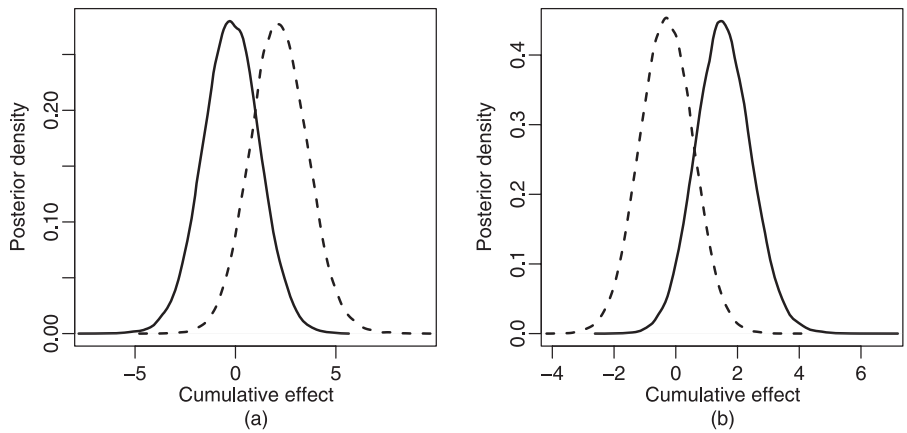

Fig. 8. Joint posterior distributions of the cumulative effects for the north (———) and south (- - - - - - -) regions for (a) the COPDAE and (b) the ischaemic heart disease outcomes

ence. For each posterior sample we calculated $\xi_R^{(i)} = \Sigma_{l=0}^{L-1} \theta_{R,l}^{(i)}$ and plotted the marginal posterior distributions of the north and south cumulative effects in Fig. 8. For COPDAE, the bulk of the posterior mass for the south is to the right of the posterior for the north, with a posterior probability $\mathbb{P}(\xi_{\text{south}} > \xi_{\text{north}}|\text{data}) = 0.93$. For ischaemic heart disease, the posterior distribution for the north is centred to the right of the posterior for the south, indicating a larger effect in the north with $\mathbb{P}(\xi_{\text{north}} > \xi_{\text{south}}|\text{data}) = 0.95$.

4.2. Sensitivity analysis

One concern that is raised by applying our Bayesian DLM is the possibility that placing constraints on the parameters corresponding to longer lags would somehow introduce bias in estimates of parameters corresponding to shorter lags. To investigate this concern we estimated the national average distributed lag function by using both our BHDLM and a completely unconstrained two-stage model. This unconstrained model specifies that, for a given county c,

$$\log(\mathbb{E}[Y_t^c]) = \sum_{l=0}^{13} \theta_l^c x_{t-l}^c + \text{ other predictors},$$

and national average estimates are obtained by taking a weighted average of the county-specific distributed lag functions.

Estimates for both the unconstrained model and the BHDLM are plotted in Fig. 9. We can see that, for ischaemic heart disease, the estimates at lags 0–3 for both models are very similar, after which the BHDLM estimates are all close to 0. For COPDAE the estimates for lags 0–2 are relatively close; after lag 3 the BHDLM estimates become much more smooth than the unconstrained estimates. For both outcomes it appears that imposing constraints on the

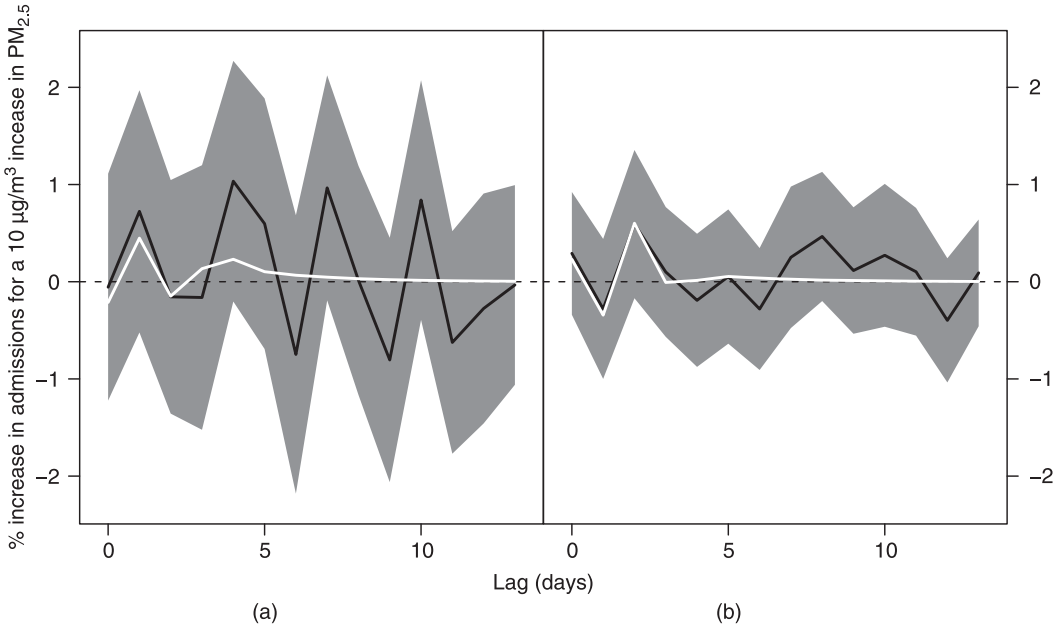

Fig. 9. Comparison of distributed lag functions estimated by the BHDLM (white curves) and by the two-stage approach using the estimated coefficients obtained from unconstrained county-specific regression models (black curves) (grey regions indicate pointwise 95% confidence intervals for two-stage model estimates): (a) COPDAE; (b) ischaemic heart disease

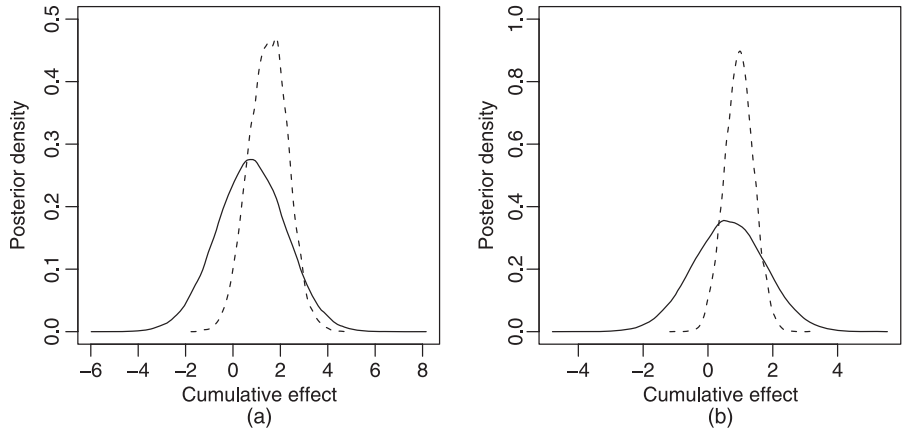

Fig. 10. Comparison of the BHDLM (———) with a model that does not allow for between-county variation (-------) for (a) COPDAE and (b) ischaemic heart disease

longer lags does not substantially bias the estimates at shorter lags in the sense that estimates at shorter lags are similar to those that would have been obtained by using an unconstrained model.

The structure of our prior distribution on the distributed lag function inherently places a 'higher standard of evidence' on the longer lags relative to the shorter lags. Since we believe *a priori* that it is more likely for there to be a strong association at lag 0 than at lag 13, the evidence for an association at lag 13 would have to be very strong for us to believe *a posteriori* that the association is non-zero. The posterior distributions of γ_1 and γ_2 and the smoothness indicated by the BHDLM model in Fig. 9 suggest that the evidence in the data for any association between $PM_{2.5}$ and hospital admissions at longer lags is very weak.

Another question that arises is to what extent allowing between-county variation in the model affects model results and inferences. As a sensitivity analysis we fit a model which does not allow for between-county heterogeneity in the distributed lag function and assumes a common function for all counties. The posterior distributions of the cumulative effects for the BHDLM and the no-heterogeneity model are plotted in Fig. 10. For both outcomes, the posterior for the cumulative effect that is obtained from the no-heterogeneity model is more narrowly dispersed than the BHDLM posterior. Given the variation in the estimates, the two models are centred at similar values, although the no-heterogeneity model is shifted slightly to the right. In general, the no-heterogeneity model and the BHDLM are centred at similar values but the no-heterogeneity model exhibits somewhat less posterior uncertainty.

5. Discussion

We have proposed a BHDLM for estimating the distributed lag between ambient air pollution levels and rates of hospitalization. The model uses a prior distribution that constrains the time course of the short-term health effects of air pollution and combines information from multiple locations. We have applied the model to a national air pollution and hospitalization database for US residents who are enrolled in Medicare, examining the relationship between $PM_{2.5}$-exposure and hospitalization for ischaemic heart disease and COPDAE.

Our model builds on the work of Welty *et al.* (2008) and Zanobetti *et al.* (2000) by smoothing distributed lag function estimates across lags and by providing a method for combining these

functions across locations where we assume more variability for parameters corresponding to shorter lags and less variability for parameters corresponding to longer lags. In addition, the hierarchical model lets us examine the range of shapes in the county-specific distributed lag functions. We have established that our methodology is related to penalized spline modelling with a special type of penalty and this connection, along with evidence from simulation studies that were conducted by Welty *et al.* (2008), creates a basis for understanding the statistical properties of our approach.

The national average distributed lag functions for COPDAE and ischaemic heart disease indicate different time courses for the effect of $PM_{2.5}$ on hospitalizations for these disease categories. The effect of $PM_{2.5}$ on COPDAE admissions appears to be spread over a longer time period than the effect on ischaemic heart disease admissions. The nature and characteristics of acute exacerbations of COPD are known to be heterogeneous across people (Sapey and Stockley, 2006) and exacerbations are often a cause of hospitalization after initial treatment outside the hospital has failed (Seemungal *et al.*, 2000). We found little evidence that the effect of an increase in $PM_{2.5}$-levels on hospitalizations for ischaemic heart disease extends beyond 2 days. In addition, the shape of the distributed lag function for ischaemic heart disease suggests some weak evidence of mortality displacement. Cardiovascular effects of PM are thought to be generally related to neurogenic and inflammatory processes (Pope *et al.*, 2003). The results from our analysis suggest that, for ischaemic heart disease in particular, the biological mechanism that is involved has a relatively short time course, with the bulk of people admitted to the hospital within 2 days of an increase in $PM_{2.5}$-levels.

When estimating the association between $PM_{2.5}$ and either health outcome, a bias–variance trade-off appears to be involved in choosing between applying a single-lag model or a DLM. Even with the national databases that were used here, estimation of the distributed lag function resulted in a substantial increase in the variance of the cumulative effect compared with risk estimates from single-lag models. Although we might consider the single-lag model's restriction to fixed lag effects a limitation (and potentially a source of bias), we must also consider the dramatic increase in precision that the model provides. If the cumulative short-term effect of an increase in air pollution levels is the sole parameter of interest, the benefits of the DLM's greater flexibility may not outweigh the cost of incurring much greater variability in the resulting estimate.

We should be careful not to overinterpret the findings of our analysis. Even with the constraints that are imposed by the prior, the uncertainty of the estimates in Fig. 3 is still large for coefficients corresponding to early lags and all the estimates have posterior intervals spanning zero. Although there appears to be substantial uncertainty in our estimation of the distributed lag function and the cumulative effect, it should be noted that only 4 years of data were available for this analysis. Previous applications of DLMs to air pollution and health data have generally had more data available. For example, the study of Schwartz (1994) made use of 8 years of data and the more recent study of Zanobetti and Schwartz (2008) examining ozone and mortality used 12 years of data. We expect that, with on-going data collection, we should be able to achieve levels of precision that are similar to previously published results.

One might question the interpretability of a national average distributed lag function in the presence of substantial variation or heterogeneity across counties. However, the usefulness of such an estimate depends ultimately on the aims of the analysis. From a regulatory standpoint, national estimates are relevant because ambient air quality standards are set at the national level and apply equally to all states. The estimates of uncertainty that are obtained by accounting for heterogeneity across counties as well as statistical variation within counties are particularly useful for policy makers. If standards were to be set at the local or regional level, then an alternative approach might need to be considered.

The north–south comparison in Section 4.1 indicated an interesting contrast between the two regions and merits further investigation. A possibility for future work could be to fit a spatial model for the distributed lag function, assuming that neighbouring counties are more similar than distant counties. The regional differences in the shapes of the distributed lag functions could be due to differences in the composition of $PM_{2.5}$ (National Research Council, 2004). The major constituents of PM pollution in the north-eastern US region include sulphate and ammonium ion, which originate largely from coal power generation sources, whereas PM in the south-eastern region generally contains more silicon, an element that is related to crustal material and mechanical processes (Bell *et al.*, 2007a). Also, the change in latitude from the north to the south covers a wide range of temperatures and climates which may alter the susceptibilities of populations to air pollution exposure.

One limitation of our application of the BHDLM is the reliance on the Poisson distribution in the county-specific model (2). Whereas previous time series studies of air pollution and health have suggested that there is relatively little overdispersion in the residuals, a more flexible alternative might be to use a generalized Poisson model (Consul and Famoye, 1992). Another point of discussion concerns the prior distributions that were used in this application. We have placed uniform hyperprior distributions on $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$ which place equal prior weight on models which may not be equally plausible. Nevertheless, the posterior distributions in Fig. 4 suggest that there is some information in the data to choose between these models.

It should be noted that the marginal posterior distribution for γ_1 is concentrated on the boundary of the allowable range for γ_1, making the interpretation of that posterior distribution potentially sensitive to the location of the boundary. As noted in Appendix A, we have constructed this range to be as wide as possible to encompass a range of distributed lag functions. It is possible that the uniform hyperprior for γ_1 is not optimal in this situation and a modification of our approach is a potential topic for further development.

The specific use of an exponential decay in the variance of the lag coefficients does affect the resulting shape of the estimated distributed lag function somewhat. We have explored alternative decay functions such as a half-normal and power law and our analyses indicate that these alternatives do not affect the substantive conclusions of the investigation.

Our model did not include any interactions between levels of $PM_{2.5}$ on different days or with averages of $PM_{2.5}$-levels over several days. It is plausible that such interactions exist and, if so, estimates from our model are likely to be biased. In our initial exploratory analyses models containing simple interactions were fitted and we generally found little evidence to support their inclusion. Nevertheless, the development of a more structured approach to the estimation of interactions as well as the development of appropriate prior distributions are an important direction for future work.

Finally, Medicare data are collected for administrative purposes and diagnoses of disease are known to be subject to some missclassification. However, such missclassification would only bias our results if the daily pattern of diagnosis and coding varied in a way that was correlated with $PM_{2.5}$-levels.

The principal benefit of the DLM is its ability to estimate the shape of the distributed lag function relating increases in air pollution to health outcomes in short periods of time after an air pollution episode. Our model provides a useful parameterization that can easily incorporate prior knowledge and be applied to large multisite databases. Over time, as more data become available from national databases, our model could be applied to track the health effects of PM. Although the results of our analysis are interesting and suggest some possible hypotheses, more focused studies (perhaps involving compositional data on $PM_{2.5}$ or susceptible subpopulations)

will have to be conducted to obtain more precise information about the biological mechanisms that are involved.

Although we have focused here on the application of our model to air pollution and health data, our model should be applicable to a wide range of substantive areas where the effects of an exposure or input on an outcome is spread over multiple time points. DLMs have already been used extensively in economics where the effect of certain indicators on economic outputs of interest often have a distributed lag structure. Although past uses of DLMs have been in the single-location setting (e.g. Almon (1965) and Leamer (1972)), the increasing availability of data from around the USA and the world increases the relevance of our hierarchical model. In addition to economics, our BHDLM should be applicable more generally in the area of environmental statistics. For example, a growing area of research involves the health impacts of climate change and the effects of temperature on mortality and morbidity (e.g. Bell *et al.* (2007b)). Since temperature data are much more abundant than air pollution data, application of our BHDLM here seems potentially fruitful. In general, with the increasing sophistication of data collection systems in many scientific areas providing more data from multiple locations, the applicability and relevance of our model will probably increase.

Acknowledgements

This research was supported in part by a Faculty Innovation Fund award from the Johns Hopkins Bloomberg School of Public Health, grant RD-83241701 from the US Environmental Protection Agency, the National Institute of Environmental Health Sciences Center in Urban Environmental Health (grant P30ES03819) and grant ES012054-03 from the National Institute of Environmental Health Sciences.

Appendix A: Details of Gibbs sampler

We implement a hybrid Gibbs sampler to sample from the posterior distributions of $\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\theta}_c$ $(c = 1, \ldots, n)$ and $\boldsymbol{\mu}$. Briefly, the full conditionals for $\boldsymbol{\eta}, \boldsymbol{\gamma}$ and $\boldsymbol{\theta}_c$ for $c = 1, \ldots, n$ are sampled by using a Metropolis–Hastings rejection step and the full conditional for $\boldsymbol{\mu}$ is sampled in closed form. All calculations were done using R version 2.4.1 (R Development Core Team, 2006). We describe the procedures for sampling from the full conditional distributions below.

(a) *Sampling* $\boldsymbol{\theta}_c$: to sample from the full conditional for $\boldsymbol{\theta}_c$ we implement a Metropolis–Hastings rejection scheme. Sampling from the full conditional for $\boldsymbol{\theta}_c$ requires evaluting the likelihood for county c with both $\boldsymbol{\theta}_c$ and the nuisance parameters in $\boldsymbol{\beta}_c$. Rather than assume a prior distribution for the many nuisance parameters in $\boldsymbol{\beta}_c$, we evaluate the profile likelihood $L_p(\boldsymbol{\theta}_c) = \max_{\beta_c}\{L_f(\boldsymbol{\theta}_c, \boldsymbol{\beta}_c)\}$, where, for each given value of $\boldsymbol{\theta}_c$, we maximize the full Poisson likelihood L_f with respect to $\boldsymbol{\beta}_c$, holding $\boldsymbol{\theta}_c$ fixed. In the Metropolis–Hastings step taken to sample from the full conditional for $\boldsymbol{\theta}_c$, we use the profile likelihood for $\boldsymbol{\theta}_c$ to calculate the acceptance ratio for the proposal. The proposal distribution for sampling from the full conditional of $\boldsymbol{\theta}_c$ is constructed by first estimating $\boldsymbol{\theta}_c$ in a county-specific log-linear Poisson regression model to obtain $\hat{\boldsymbol{\theta}}_c$ and its estimated covariance matrix $\hat{\Sigma}_c$. If we assume as in the two-stage approach that $\hat{\boldsymbol{\theta}}_c|\boldsymbol{\theta}_c \sim \mathcal{N}(\boldsymbol{\theta}_c, \hat{\Sigma}_c)$, we can compute the conditional distribution of $\boldsymbol{\theta}_c$ given $\hat{\boldsymbol{\theta}}_c$ and the current values of $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ and use this conditional distribution as a proposal distribution, i.e.

$$\boldsymbol{\theta}_c^* | \hat{\boldsymbol{\theta}}_c, \boldsymbol{\mu}, \boldsymbol{\gamma} \sim \mathcal{N}\{\boldsymbol{\mu} + B_1(\hat{\boldsymbol{\theta}}_c - \boldsymbol{\mu}), \sigma_\gamma^2 (I - B_1)\, \Omega(\boldsymbol{\gamma})\} \tag{11}$$

where $B_1 = \sigma_\gamma^2\, \Omega(\boldsymbol{\gamma})\{\hat{\Sigma}_c + \sigma_\gamma^2\, \Omega(\boldsymbol{\gamma})\}^{-1}$. Given the proposal distribution in expression (11), the full conditional for $\boldsymbol{\theta}_c$ is then proportional to

$$p(\boldsymbol{\theta}_c|\cdot) \propto L_p(\boldsymbol{\theta}_c)\, \varphi\{\boldsymbol{\theta}_c | \boldsymbol{\mu}, \sigma_\eta^2\, \Omega(\boldsymbol{\eta})\}$$

where $\varphi\{\boldsymbol{\theta}_c|\boldsymbol{\mu}, \sigma_\eta^2\, \Omega(\boldsymbol{\eta})\}$ is the multivariate normal density with mean $\boldsymbol{\mu}$ and covariance matrix $\sigma_\eta^2\, \Omega(\boldsymbol{\eta})$ and $L_p(\boldsymbol{\theta}_c)$ is the profile likelihood for $\boldsymbol{\theta}_c$.

(b) *Sampling* $\boldsymbol{\mu}$: the full conditional for $\boldsymbol{\mu}$ is proportional to

$$p(\boldsymbol{\mu}|\cdot) \propto \Big[\prod_{c=1}^{n} \varphi\{\boldsymbol{\theta}_c|\boldsymbol{\mu}, \sigma_\eta^2\,\Omega(\boldsymbol{\eta})\}\Big]\varphi\{\boldsymbol{\mu}|\mathbf{0}, \sigma_\gamma^2\,\Omega(\boldsymbol{\gamma})\}$$
$$= \mathcal{N}\{B_2\bar{\boldsymbol{\theta}}, (I - B_2)\sigma_\gamma^2\,\Omega(\boldsymbol{\gamma})\}$$

where $B_2 = \sigma_\gamma^2\,\Omega(\boldsymbol{\gamma})\{\sigma_\gamma^2\,\Omega(\boldsymbol{\gamma}) + \sigma_\eta^2\,\Omega(\boldsymbol{\eta})/n\}^{-1}$ and $\bar{\boldsymbol{\theta}} = (1/n)\Sigma\boldsymbol{\theta}_c$.

(c) *Sampling* $\boldsymbol{\eta}$ *and* $\boldsymbol{\gamma}$: we put uniform priors on both $\boldsymbol{\eta} = (\eta_1, \eta_2)$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$ and hence the full conditionals for $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$ are

$$p(\boldsymbol{\eta}|\cdot) \propto \prod_{c=1}^{n} \varphi\{\boldsymbol{\theta}_c|\boldsymbol{\mu}, \sigma_\eta^2\,\Omega(\boldsymbol{\eta})\}$$

and

$$p(\boldsymbol{\gamma}|\cdot) \propto \varphi\{\boldsymbol{\mu}|\mathbf{0}, \sigma_\gamma^2\,\Omega(\boldsymbol{\gamma})\}.$$

To preserve numerical stability, we placed upper and lower bounds on each parameter so that both η_1 and η_2 were restricted to be in the range $[0.2, 0.8]$ whereas γ_1 and γ_2 were restricted to be in the range $[0.05, 0.75]$. These bounds were chosen on the basis of previous work and some exploratory analysis. Upper bounds that were much larger than these values often produced covariance matrices that were not invertible. We subsequently used uniform proposal distributions (restricted to the appropriate ranges) and a Metropolis–Hastings rejection step to sample from the full conditionals of $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$.

The Gibbs samplers for each hospitalization outcome were each run for 40000 iterations with 10000 iterations discarded as burn-in. Acceptance percentages for the Metropolis–Hastings steps were tuned to be between 10% and 30%. Convergence of the chains was diagnosed by estimating Monte Carlo standard errors of the parameters by using the method of batch means that was described in Jones *et al.* (2006).

References

Almon, S. (1965) The distributed lag between capital appropriations and expenditures. *Econometrica*, **33**, 178–196.

Bell, M. and Davis, D. (2001) Reassessment of the lethal London fog of 1952: novel indicators of acute and chronic consequences of acute exposure to air pollution. *Environ. Hlth Perspect.*, **109**, 389–394.

Bell, M. L., Dominici, F., Ebisu, K., Zeger, S. L. and Samet, J. M. (2007a) Spatial and temporal variation in $PM_{2.5}$ chemical composition in the United States for health effects studies. *Environ. Hlth Perspect.*, **115**, 989–995.

Bell, M. L., Goldberg, R., Hogrefe, C., Kinney, P. L., Knowlton, K., Lynn, B., Rosenthal, J., Rosenzweig, C. and Patz, J. (2007b) Climate change, ambient ozone, and health in 50 U.S. cities. *Clim. Change*, **82**, 61–76.

Bell, M. L., McDermott, A., Zeger, S. L., Samet, J. M. and Dominici, F. (2004) Ozone and short-term mortality in 95 US urban communities, 1987-2000. *J. Am. Med. Ass.*, **292**, 2372–2378.

Chatfield, C. (1996) *The Analysis of Times Series: an Introduction*, 5th edn. London: Chapman and Hall.

Clarke, R. W., Catalano, P. J., Koutrakis, P., Murthy, G. G., Sioutas, C., Paulauskis, J., Coull, B., Ferguson, S. and Godleski, J. J. (1999) Urban air particulate inhalation alters pulmonary function and induces pulmonary inflammation in a rodent model of chronic bronchitis. *Inhaln Toxicol.*, **11**, 637–656.

Consul, P. C. and Famoye, F. (1992) Generalized Poisson regression model. *Communs Statist. Theory Meth.*, **21**, 89–109.

Corradi, C. (1977) Smooth distributed lag estimators and smoothing spline functions in Hilbert spaces. *J. Econometr.*, **5**, 211–220.

Dominici, F., Daniels, M., Zeger, S. L. and Samet, J. M. (2002a) Air pollution and mortality: estimating regional and national dose-response relationships. *J. Am. Statist. Ass.*, **97**, 100–111.

Dominici, F., McDermott, A., Zeger, S. L. and Samet, J. M. (2002b) Airborne particulate matter and mortality: time-scale effects in four US Cities. *Am. J. Epidem.*, **157**, 1053–1063.

Dominici, F., McDermott, A., Zeger, S. L. and Samet, J. M. (2003) Airborne particulate matter and mortality: timescale effects in four US cities. *Am. J. Epidem.*, **157**, 1055–1065.

Dominici, F., Peng, R. D., Bell, M. L., Pham, L., McDermott, A., Zeger, S. L. and Samet, J. M. (2006) Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *J. Am. Med. Ass.*, **295**, 1127–1134.

Everson, P. J. and Morris, C. N. (2000) Inference for multivariate normal hierarchical models. *J. R. Statist. Soc.* B, **62**, 399–412.

Goodman, P. G., Dockery, D. W. and Clancy, L. (2004) Cause-specific mortality and the extended effects of particulate pollution and temperature exposure. *Environ. Hlth Perspect.*, **112**, 179–185.

Health Effects Institute (2003) Revised analyses of time-series studies of air pollution and health. *Special Report*. Health Effects Institute, Boston.

Huang, Y., Dominici, F. and Bell, M. L. (2005) Bayesian hierarchical distributed lag models for summer ozone exposure and cardio-respiratory mortality. *Environmetrics*, **16**, 547–562.

Jones, G. L., Haran, M., Caffo, B. S. and Neath, R. (2006) Fixed-width output analysis for Markov Chain Monte Carlo. *J. Am. Statist. Ass.*, **101**, 1537–1547.

Katsouyanni, K., Toulomi, G., Samoli, E., Gryparis, A., LeTertre, A., Monopolis, Y., Rossi, G., Zmirou, D., Ballester, F., Boumghar, A. and Anderson, H. R. (2001) Confounding and effect modification in the short-term effects of ambient particles on total mortality: results from 29 European cities within the APHEA2 Project. *Epidemiology*, **12**, 521–531.

Kelsall, J. E., Samet, J. M., Zeger, S. L. and Xu, J. (1997) Air pollution and mortality in Philadelphia, 1974–1988. *Am. J. Epidem.*, **146**, 750–762.

Lay, J. C., Bennett, W. D., Ghio, A. J., Bromberg, P. A., Costa, D. L., Kim, C. S., Koren, H. S. and Devlin, R. B. (1999) Cellular and biochemical response of the human lung after intrapulmonary instillation of ferric oxide particles. *Am. J. Resp. Cell Molec. Biol.*, **20**, 631–642.

Leamer, E. E. (1972) A class of informative priors and distributed lag analysis. *Econometrica*, **40**, 1059–1081.

Lippmann, M., Ito, K., Hwang, J.-S., Maciejczyk, P. and Chen, L.-C. (2006) Cardiovascular effects of nickel in ambient air. *Environ. Hlth Perspect.*, **114**, 1662–1669.

National Research Council (2004) *Research Priorities for Airborne Particulate Matter*, vol. IV, *Continuing Research Progress*. National Research Council of the National Academies.

Peng, R. D., Dominici, F. and Louis, T. A. (2006) Model choice in time series studies of air pollution and mortality (with comments). *J. R. Statist. Soc.* A, **169**, 179–203.

Peng, R. D., Dominici, F., Pastor-Barriuso, R., Zeger, S. L. and Samet, J. M. (2005) Seasonal analyses of air pollution and mortality in 100 US cities. *Am. J. Epidem.*, **161**, 585–594.

Pope, C. A., Burnett, R. T., Thruston, G. D., Calle, E., Thun, M. J., Krewski, D. and Goldeski, J. (2003) Cardiovascular mortality and long-term exposure to particulate air pollution: epidemiological evidence of general pathophysiological pathways of disease. *Circulation*, **6**, 71–77.

Pope, C. A. and Dockery, D. W. (2006) Health effects of fine particulate air pollution: lines that connect. *J. Air Wste Mangmnt Ass.*, **56**, 709–742.

R Development Core Team (2006) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Roberts, S. (2005) An investigation of distributed lag models in the context of air pollution and mortality time series analysis. *J. Air Wste Mangmnt Ass.*, **55**, 273–282.

Samoli, E., Touloumi, G., Zanobetti, A., Le Tertre, A., Schindler, C., Atkinson, R., Vonk, J., Rossi, G., Saez, M., Rabczenko, D., Schwartz, J. and Katsouyanni, K. (2003) Investigating the dose-response relation between air pollution and total mortality in the APHEA-2 multicity project. *Occupnl Environ Med.*, **60**, 977–982.

Sapey, E. and Stockley, R. A. (2006) COPD exacerbations 2: aetiology. *Thorax*, **61**, 250–258.

Schimmel, H. and Murawski, T. J. (1976) The relation of air pollution to mortality. *J. Occupnl Med.*, **18**, 316–333.

Schmidt, A. M., de Fátima da G. Conceição, M. and Morerira, G. A. (2007) Investigating the sensitivity of Gaussian processes to the choice of their correlation function and prior specifications. *J. Statist. Computn Simuln*, to be published.

Schwartz, J. (1994) Nonparametric smoothing in the analysis of air pollution and respiratory illness. *Can. J. Statist.*, **22**, 471–488.

Schwartz, J. (2000) The distributed lag between air pollution and daily deaths. *Epidemiology*, **11**, 320–326.

Seemungal, T. A. R., Donaldson, G. C., Bhowmik, A., Jeffries, D. J. and Wedzicha, J. A. (2000) Time course and recovery of exacerbations in patients with chronic obstructive pulmonary disease. *Am. J. Resp. Crit. Care Med.*, **161**, 1608–1613.

Shiller, R. J. (1973) A distributed lag estimator derived from smoothness priors. *Econometrica*, **41**, 775–788.

Welty, L. J., Peng, R. D., Zeger, S. L. and Dominici, F. (2008) Bayesian distributed lag models: estimating the effects of particulate matter air pollution on daily mortality. *Biometrics*, to be published.

Welty, L. J. and Zeger, S. L. (2005) Are the acute effects of PM_{10} on mortality in NMMAPS the result of inadequate control for weather and season?: a sensitivity analysis using flexible distributed lag models. *Am. J. Epidem.*, **162**, 80–88.

Zanobetti, A. and Schwartz, J. (2008) Mortality displacement in the association of ozone with mortality. *Am. J. Resp. Crit. Care Med.*, **177**, 184–189.

Zanobetti, A., Schwartz, J., Samoli, E., Gryparis, A., Touloumi, G., Atkinson, R., Le Tertre, A., Bobros, J., Celko, M., Goren, A., Forsberg, B., Michelozzi, P., Rabczenko, D., Aranguez, R. E. and Katsouyanni, K. (2002) The temporal pattern of mortality responses to air pollution: a multicity assessment of mortality displacement. *Epidemiology*, **13**, 87–93.

Zanobetti, A., Wand, M., Schwartz, J. and Ryan, L. (2000) Generalized additive distributed lag models: quantifying mortality displacement. *Biostatistics*, **1**, 279–292.

Zeger, S. L., Dominici, F. and Samet, J. M. (1999) Harvesting-resistant estimates of pollution effects on mortality. *Epidemiology*, **89**, 171–175.